DOCUMENT RESUME

ED 453 656 FL 026 716

AUTHOR Murray, Joel R.
TITLE Steps and Recommendations for More Accurate Placement Test Creation.
PUB DATE 2001-05-00
NOTE 50p.
PUB TYPE Information Analyses (070) -- Reports - Evaluative (142)
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *English (Second Language); Examiners; Factor Analysis; Literature Reviews; Second Language Instruction; Second Language Learning; *Student Evaluation; *Student Placement; Teacher Education; *Test Construction; Test Format; Test Theory; Test Validity

ABSTRACT

This paper aims to provide practical advice for creating a placement test for English-as-a-Second-Language (ESL) or English-as-a-foreign-language (EFL) instruction. Three forms of concrete assistance are provided: a detailed literature review; detailed steps focusing on the creation of placement tests; and a set of recommendations focusing on individual parts of placement tests. The literature review defines a placement test, explores the components of the test in methodical detail covering listening comprehension, speaking, grammar, vocabulary, reading, writing, and language test creation, methodology, and practice. Recommended steps to take in the creation of a useful and accurate placement test include assembling an assessment team, defining test takers and objectives, developing rubrics and rating scales, testing the test itself, and training the scorers and administrators of the test. Recommendations focusing on individual parts of the placement test cover listening, reading, speaking, writing, and grammar. (Contains 28 references.) (KFT)

STEPS AND RECOMMENDATIONS FOR MORE ACCURATE PLACEMENT TEST CREATION

Joel R. Murray

TABLE OF CONTENTS

INTRODUCTION 1

A REVIEW OF THE LITERATURE 3

Testing And Assessment 4
What is a Placement Test? 4
The Components of Second Language Placement Tests 5
Tests of Listening Comprehension 5
Tests of Speaking 7
Tests of Grammar 8
Tests of Vocabulary 8
Tests of Reading 9
Tests of Writing 11
Language Test Creation, Methodology, and Practice 12
Steps and Recommendations in Placement Test Creation 13
Other Considerations 15
Issues Related to Accurate Placement 15
The Ethics and Effects of Testing 22

STEPS TO TAKE IN THE CREATION OF A USEFUL AND ACCURATE PLACEMENT TEST 25

Assembling an Assessment Team 25
Defining Characteristics of the Test-takers 26
Defining Objectives for the Placement Test 27
Deciding on the Type of Test to be Used and Its Contents 27
Creating the Test 30
Developing Rubrics or Rating Scales for the Test 31
Testing the Placement Test Itself 33
Training the Scorers and Administrators of the Test 34

RECOMMENDATIONS FOCUSING ON INDIVIDUAL PARTS OF THE PLACEMENT TEST 36

Listening 36
Reading 37
Speaking 39
Writing 41
Grammar 42

CONCLUSION 43

REFERENCES 44

INTRODUCTION

To paraphrase Elson (1992), there has been much diverse work on issues related to the testing of English as a second language (ESL) or English as a foreign language (EFL) within the past 25 years. One issue that has received attention has been the topic of placement testing, the initial testing most students undergo upon entering an ESL or EFL program. For the most part, however, placement testing has been approached from the perspective of issues surrounding the test itself, with some authors, for example, describing the implementation of a placement test (Chandavimol, 1988; Malu, 1989) or others, noting problems with a particular placement test, suggesting either methods to increase the accuracy of the test (Brown, 1989; Ilyin, 1970; Rich, 1993) or the use of alternative forms of testing (LeBlanc & Painchaud, 1985). Although much has been written concerning such issues, relatively little has been written regarding the actual creation of placement tests. In other words, relatively little practical help exists for those educators tasked with creating a placement test. Thus, this article will offer three forms of assistance: first, a detailed literature review; second, detailed steps focusing on the creation of placement tests; and third, a set of recommendations focusing on individual parts of placement tests.

A key question at this point is why there may be a need for this type of assistance. One reason is that, at many institutes, placement testing has evolved slowly and almost haphazardly in response to external pressures—for example, an increasing capacity for students or a widening selection of classes into which to place those students. The result is that often there is no clear-cut plan for placement testing, and those tasked with producing such a test invariably have little training in placement test creation. Therefore, the accuracy of a placement instrument created

under these conditions is bound to be questionable. In these situations, then, a set of steps and recommendations would be invaluable.

A second reason is that while guides containing steps and recommendations for placement testing do exist, they tend to be somewhat outdated (Harrison, 1983) or mentioned briefly as part of an overall discussion of other matters (Brown, 1995; Hughes, 1989). Another reason is that steps and specifications for the creation of any test are "a central and crucial part of the test construction and evaluation process" (Alderson, Clapham, & Wall, 1995, p. 9), and that they are needed by a wide variety of people, such as the constructors of tests, the users of tests, the test-takers, teachers, administrators, students, and those responsible for establishing test validity (Alderson, et al., 1995).

A final reason is that both the literature review and the steps and recommendations presented in this article should offer to those involved in devising placement tests a much needed systematic or rational basis for developing their tests—in other words, a guideline, or what Lynch and Davidson (1994) refer to as a "blueprint" that test writers and test administrators can use in the creation and administration of their placement tests. As Brown (1995) points out, "though all this may seem like a great deal of work, remember that in most language programs, any rational approach to testing will be a vast improvement over the existing conditions" (p. 119) and "the work is worthwhile because of the information that can be gained and the satisfaction that can be derived from making responsible decisions about students' lives" (p. 123).

A REVIEW OF THE LITERATURE

This literature review touches on several issues concerning placement testing and will be divided into two main sections: "Testing and Assessment" and "Important Considerations." In "Testing and Assessment," the four types of test—achievement, proficiency, diagnostic, and placement—will be discussed first in order (a) to establish that placement tests are distinct entities, different from other types of test, and (b) to define the term "placement test." The components of language tests—tests of listening, speaking, grammar, vocabulary, writing, and reading—will be examined next, as all or a combination of these components are found in placement tests. These components will be discussed in terms of important considerations and implications in reference to testing generally and placement testing specifically. Language test creation, methodology, and practice will be discussed next, for these points are crucial in the analysis of any test. Finally, steps and recommendations in test creation will be reviewed, as there already exist suggestions or advice that may be applicable to the creation of an accurate placement vehicle.

In "Important Considerations," the significance of accurate placement testing and the steps that were taken to improve placements in various environments will be examined. Accuracy in placement testing is important, and a variety of researchers have been faced with placement tests that have appeared inadequate to the task of placing test-takers into appropriate classes. The ethics and effects of testing will also be examined, as any type of testing can be considered to be powerful, having far-reaching effects on the lives of test-takers.

Testing And Assessment

What is a Placement Test?

Many authors (Alderson, et al., 1995; Harrison, 1983; Hughes, 1989; Underhill, 1991) agree that there are four types of test: achievement, proficiency, diagnostic, and placement. Achievement tests are those which are given in order to assess what the students in a particular course of study have learned. Some authors (Alderson, et al.; Hughes) recognize two kinds of achievement tests: progress, which are administered at various stages throughout a language course, and final, which are administered at the end of a course. Proficiency tests are those which are given in order to measure general language ability regardless of previous language training. They are different from achievement tests in that proficiency tests are not based on a particular language program. Instead, they are based on "a specification of what candidates have to be able to do in the language in order to be considered proficient" (Hughes, 1989, p. 9). Diagnostic tests are those which are given in order to determine the test-takers' strengths and weaknesses. Used to identify areas in which the test-taker needs help, few purely diagnostic tests exist "since it is difficult to diagnose precisely strengths and weaknesses in the complexities of language ability" (Alderson, et al, p. 12). For this reason, achievement and proficiency tests are often used for this purpose.

Placement tests are seen to be different from the other three types of tests. Harrison (1983) defines "placement test" as a test that is created so as "to sort new students into teaching groups, so that they can start a course at approximately the same level as the other students in the class" (p. 4). He adds that placement tests are concerned with the test-taker's present state of general language ability rather than with "specific points of learning" (p. 4), and as a result, "a variety of tests is necessary because a range of different activities is more likely to give an

accurate overall picture of a student's level than a single assessment" (p. 4). While accurate in a general sense, Harrison's definition may be somewhat limited, in that not every placement test may be one of *overall* language ability. By way of illustration, a language program that has a dominant oral/aural focus is unlikely to employ a placement test that places a great deal of emphasis on reading comprehension or writing. Hughes (1989) provides a definition similar to that of Harrison but different in that Hughes' is not as narrow. He states that placement tests "provide information which will help to place students at the stage (or in the part) of the teaching programme most appropriate to their abilities. Typically they are used to assign students to classes at different levels" (p. 14). Alderson, et al. (1995) provide one of the best definitions. They state that "placement tests are designed to assess students' level of language ability so that they can be placed in the appropriate course or class. Such tests may be based on aspects of the syllabus taught at the institution concerned, or may be based on unrelated material" (p. 11).

The Components of Second Language Placement Tests

An examination of the components of second language placement tests follows. These components are explored in terms not of what they are, what should be tested, and how, but of important considerations and implications in reference to testing generally and placement testing specifically.

Tests of Listening Comprehension

Tests of listening comprehension seek to "assess the ability to use knowledge of the language for the purpose of understanding spoken texts" (Buck, 1997, p. 71). In reviewing and detailing the problems inherent in, history of, and practical advice about the testing of listening

in a second language, Buck makes a number of important points. First, he asserts that testing listening comprehension is by necessity indirect; thus, "listening scores will always be influenced by other skills required for task completion" (p. 66). This point is important in an overall sense in that listening tests may not assess listening alone, and as a result, accurate assessments of listening comprehension may not be possible. Second, he points out that listeners' interpretation of any spoken text will be influenced to a large degree by their purposes for listening in the first place, their interests, and their background knowledge. There can be, therefore, a variety of valid interpretations of a text, some of which cannot be anticipated. Third, he states that "virtually all second language listening tests use non-interactive tasks, that is tasks in which the listener cannot interact with the speaker; interactive listening is usually only assessed as part of a spoken interview" (p. 66). Buck adds that "traditionally testers have not been interested in visual media for the presentation of listening texts" (p. 72), the result of which has been a removal of what Weir (1990) refers to as "the wealth of normal exophoric reference and paralinguistic information" (p. 54) found in the visual element.

On the topic of assessing listening comprehension through the use of visual media, Progosh (1996) reports on the value of using video in conjunction with listening assessment. He conducted a study to determine test-takers' opinions of a video-mediated listening test by using a random sample of the second year population of intermediate-level second language students at a tertiary institution in Tokyo, Japan. Using a questionnaire consisting mainly of questions answered on a seven-point Likert scale, Progosh found that "the sample think [sic] video in listening comprehension is a good idea, preferring video-mediated tests over audiocassette tests" (p. 40). He does warn, however, that the video in this case was used to assess learner achievement and that it has "yet to be determined if such tests can be used for purposes of

general language proficiency" (p. 40). Nonetheless, the use of video in conjunction with a listening assessment could be valuable in contributing to more accurate placement instruments, for as Progosh points out, "most people both hear and see in most communicative situations" (p. 35).

Tests of Speaking

Tests of speaking—or as Underhill (1991) puts it, "oral tests"—seek to assess the ability to communicate orally. Underhill states that oral tests are repeatable procedures "in which a learner speaks, and is assessed on the basis of what he says" (p. 7). This assessment is often viewed in terms of "providing information about a person which [is to] be used to predict success in communication in some future real-life situation" (Fulcher, 1997, p. 75). Fulcher, in his review of the problems inherent in, history of, and practical advice about the testing of speaking in a second language, brings up some significant issues. One is that of task; as Fulcher states, "it has been increasingly observed that task type has a systematic effect on speaking test scores" (p. 79). However, "neither the nature nor the degree of the effect of tasks on scores from tests of speaking are well understood" (p. 80). This point is notable in that importance of task type is either not mentioned in some of the literature focusing on testing spoken language (Underhill; Weir, 1990) or only mentioned in passing (Hughes, 1989), yet it may play a large role in how a test-taker's oral abilities are assessed.

On a more practical note, Underhill (1991), in his guide to oral testing, does not assume that the reader has any knowledge of language testing. Stating that the book "deplores the cult of the language testing expert" (p. 1), Underhill writes that it was written for language teachers, sequenced in the order that a test program might be implemented: starting with questions about

needs and resources, continuing with a choice of different oral test types and tasks, and discussing the marking system and evaluation. Underhill's text is useful and valuable in that it details various oral testing techniques and how to create, administer, mark, and evaluate tests of speaking. If there is to be a criticism of this text, it would be one of oversight—that is, Underhill neglects to mention that task type can be a factor on speaking test scores.

Tests of Grammar

Tests of grammar seek to assess "grammatical ability, or rather the lack of it, [for it] sets limits to what can be achieved in the way of skills performance" (Hughes, 1989, p. 142). Hughes asks whether separate grammar testing is justified in these days of communicative language testing. He states that "there is often good cause to include a grammar component in the achievement, placement and diagnostic tests of teaching institutions. It seems unlikely that there are many institutions, however 'communicative' their approach, that do not teach some grammar in some guise or other" (p. 142). As Hughes points out, "there appears to be room for a grammar component in at least some placement tests" (p. 142). Rea-Dickins adds that "the construct of grammar itself carries different meaning but is still considered by many to be an important aspect in the measurement of an individual's overall performance in a language" (p. 87).

Tests of Vocabulary

Tests of vocabulary seek to assess the knowledge of vocabulary, mostly in terms of either depth of word knowledge or size of lexicon. Read (1997) examines the assessment of vocabulary and notes that while there is a revival of interest in the teaching and learning of

vocabulary, "that has not yet led to a re-definition of the role of vocabulary within language testing or to the development of many innovative procedures for lexical assessment" (p. 99). Also on the same topic, Hughes (1989) asks whether separate vocabulary testing is justified. He concludes that "the arguments for a separate component in other kinds of test [other than proficiency tests] may not carry the same strength" (p. 147). For placement tests, Hughes suggests that "we would not normally require, or expect, a particular set of lexical items to be a prerequisite for a particular language class. All we would be looking for is some general indication of the adequacy of the student's vocabulary" (p. 147). The problem is, of course, how to go about doing so.

Tests of Reading

Tests of reading seek to assess the ability of a reader to "extract an agreed level of meaning under specified performance conditions" (Weir, 1997, p. 39). In exploring the testing of reading in a second language, Weir reviews early developments, test methods, test validation methods, work in progress regarding the construct of reading, and the problems inherent in testing reading. While noting that "a direct reading test should reflect as closely as possible the interaction that takes place between a reader and a text in the equivalent real life reading activity" (p. 39), Weir admits that "although full genuineness of text or authenticity of task is likely to be unattainable in the second language reading tests we develop, we still need to select appropriate texts, to be read for realistic purposes, and we expect the reader to extract an agreed level of meaning under specified performance conditions" (p. 39). Hughes (1989) has some very practical advice concerning the testing of reading comprehension. He suggests specifying, as accurately and completely as possible, the abilities to be tested, and he discusses the selection of

test content and the setting of criterial levels of performance. Both Weir and Hughes are important as they both address the advisability of authenticity (or something that approaches it) in a test of reading comprehension.

Courchêne (1995), also on a practical note, offers an alternative to the multiple-choice tests often used in testing reading comprehension: the summary cloze technique. A summary cloze is prepared by summarizing the content of a text so that the resulting new text is approximately one-third the length of the original. This shorter version is transformed into a cloze using a "rational deletion approach as opposed to deleting every nth word" (p. 52). Although he allows that the use of the cloze test has been challenged as a measure of language proficiency, Courchêne maintains that "the cloze procedure . . . can be used to measure reading comprehension if one selects texts of general interest to students, uses a rational as opposed to a random deletion of items, pretests them on both native and nonnative speakers, and uses them in a foreign language context" (p. 51).

In order to demonstrate that the summary cloze technique is at least as good a measure of reading comprehension as the multiple-choice format often used in testing reading comprehension, Courchêne tells of a study using 66 Chinese students at intermediate and advanced levels who were to come to Canada for academic and professional reasons. They were matched for language ability and then randomly assigned to one of two groups. These groups were given five reading passages which were controlled for length, difficulty, and reading level, and which were prepared with both multiple-choice questions and summary cloze formats. Each group did two texts in one format and two in the other, and both did the summary cloze and multiple-choice. The results were compared, and the summary cloze was correlated with other measures of language proficiency. Courchêne found that the summary cloze technique

"produces tests that tend to yield higher levels of reliability than their [multiple-choice] counterparts" (p. 56). Furthermore, he found that "the correlations of the task types with general measure of ESL proficiency do provide evidence that there are no substantial differences in the way the tasks behave, and in general the assumption holds that the two task types are both measures of reading comprehension" (p. 57). There is at least one problem with this new technique: the choice of text and its summarization may affect the test-takers ability to respond to it. Nevertheless, Courchêne's article is important in that he offers a viable alternative to the common multiple-choice reading comprehension sections of most placement tests. As Courchêne states, "initial use in the classroom and as a testing instrument have resulted in positive feedback from the students [who] see summary cloze as having face validity" (p. 57).

Tests of Writing

Tests of writing seek to assess the ability in writing "to make effective use of varied, complex aspects of language proficiency in a purposeful manner [while] providing, for the purposes of assessment, direct evidence of individual students' language performance" (Cumming, 1997, p. 51). Weir (1990) looks at the testing of writing and comments that two different approaches can be taken: the assessment of writing can be indirect and divided into discrete levels (such as grammar, vocabulary, spelling, and punctuation) and tested objectively, and it can be direct, through extended writings tasks, and tested subjectively. He suggests that writing be tested directly, and maintains that "the writing component of any test should concentrate on controlled writing tasks where features of audience, medium, setting and purpose can be more clearly specified" (p. 73). Weir also discusses the great importance of reliable marking schemes; he compares two approaches to marking, analytical and general impression;

looks at multiple marking; reviews holistic scoring; and considers factors which may contribute to the reliability of a writing test. Hughes (1989) agrees with Weir in stating that "the best way to test people's writing ability is to get them to write" (p. 75). He also notes the importance of reliable scoring and looks at the analytical and holistic approaches to marking. Weir and Hughes are important in that they both advocate the direct testing of writing and place a great amount of emphasis on the reliability of marking schemes.

Language Test Creation, Methodology, and Practice

Up to this point, the four types of test have been examined briefly, the placement test has been defined, and the components of a second language test have been surveyed. Language test creation, methodology, and practice will now be discussed, as these points are crucial in the analysis of any language test.

Weir (1990) looks at the implications of the communicative approach in terms of language testing and examines discrete point, integrative, and communicative approaches to language testing. He details such terms important to any discussion of testing as reliability, validity, and efficiency, and makes a point about the concept of face validity, an important concept in the discussion of any type of test. Weir's text is important for two reasons. First, he covers the design, development, operation, and monitoring of tests, all of which are important in the analysis of testing and in the creation of specifications for accurate placement tests. Second, he makes a strong case for communicative testing by reviewing the deficiencies of discrete-point and integrated testing.

Hughes (1989) looks at testing from a language teacher's perspective. Like Weir (1990), he reviews terms and offers suggestions on the design and development of tests. Although

looking at testing from the view of a language teacher would seem at first glance to be incompatible with the purpose of this thesis, Hughes does make some valid observations. First, he notes that "very often, [tests] fail to measure accurately whatever it is that they are intended to measure. Teachers know this. Students' true abilities are not always reflected in the test scores that they obtain" (p. 2). Second, he identifies test content and testing techniques to be sources of inaccuracy. Third, he reviews test techniques for testing overall ability, which placement testing seeks to assess. Finally, Hughes offers some practical advice on how a placement test might be designed. Unfortunately, although his advice contains ideas that are of value to anyone endeavoring to create a language test (his stages of test construction are particularly useful), his example of general procedures for the construction of a placement test "for a commercial English language teaching institution" (p. 55) contains recommendations for format (cloze tests and partial dictations) that alone could hardly contribute to accurate placement.

Steps and Recommendations in Placement Test Creation

Steps and recommendations in placement test creation will now be discussed, as there already exist suggestions or advice that may be applicable to the creation of an accurate placement vehicle.

Hughes (1989) looks at test creation from the teacher's perspective and discusses important concepts in testing, such as the kinds of tests and testing, marking, and validity and reliability. Hughes' text is particularly valuable in that he offers practical suggestions on testing writing, oral ability, reading, listening, and grammar and vocabulary, and couples these with useful examples. Harrison (1983) also looks at test creation from the teacher's point of view and explains basic principles and concepts in testing, different types of tests, marking, and

procedures to help interpret scores and the efficiency of the test itself. While his book is now a little dated and more a practical survey of different types of tests (placement, diagnostic, achievement, proficiency) and things related to them (marking, statistics, and so on), it is useful in that it offers a valuable list of specifications for a placement test, and a commentary on them. This list, which contains information on objectives of the placement test, the skills to be tested, the content of the test, its format, rubrics to be used, materials, and marking, could be used as a guide in the construction of a list of recommendations for other language institutes or programs to follow.

Carroll (1980) aims to "outline principles and techniques for specifying the communicative needs of a language learner and for assessing his language performance in terms of those needs" (p. 5). In order to do so, he gives suggestions on the design and development of communicative tests and their operation, suggestions which are useful because they can then be worked into a set of steps and recommendations that one might use in creating a more accurate placement test. His text is useful for another reason; he suggests, in the design phase, describing the test-takers and analyzing their communicative needs, and he offers advice on how to go about doing so. This description and analysis of needs should prove to be helpful in the creation of an accurate placement instrument.

Other Considerations

Issues Related to Accurate Placement

Accuracy in placement testing is important, and a variety of researchers have been faced with placement tests that have appeared inadequate to the task of placing test-takers appropriately. For example, Chandavimol (1988) describes the implementation of a placement test at Chulalongkorn University in Thailand and in doing so, makes some important points. First, the author emphasizes that "accurate placement is essential" (p. 3), as the result of what the author refers to as "misclassification" (p. 3) can be student-related (e.g., misplacements and their consequences) or program-related (e.g., inadequate numbers of instructors). Second, Chandavimol looks at the testing instrument in terms of its ease of administration. The author states that "the efficiency, or the lack of it, of the test greatly depends on administrability, which depends, in turn, on a number of factors" (p. 4), one such example being the clarity of the test's instructions: "the directions must be presented in simple, uncomplicated and unequivocal language that all examinees can understand" (p. 4). In other words, miscomprehension or lack of comprehension of directions on behalf of the test-takers could lead to their not performing to the best of their abilities on the placement test, perhaps resulting in misplacements. Finally, Chandavimol points out that the quality of assessments arising from placement tests "directly depends on the time and the human resources an institution can commit to the task of grading and double-checking" (p. 4), and, although not mentioned by the author but equally as important, to the creation and implementation of an accurate placement instrument.

Further on the issue of the importance of accurate placement of students in class levels, Brown (1989) tells of his noticing that students who were placed into existing classes were different in level from those who had been promoted from lower level courses. Noting that the

type of test most often used in placement testing, the norm-referenced test (NRT), "may not necessarily measure what is being taught and learned in the courses" (p. 73), Brown outlines what he calls "a completely new strategy for constructing language placement tests" (p. 73): the combination of the useful characteristics of a criterion-referenced test (CRT) with those of a NRT "to create placement tests that not only spread students out along a continuum of language abilities (NRT), but do so on the basis of items that are demonstrably related to what the students learn while in the program (CRT)" (p. 73). After reviewing the development of a reading placement test that was based on the new strategy and intended to replace the existing test, Brown examines the item and descriptive statistics of the two forms of the test and seeks to discover how reliable the two tests were and to what degree they were valid as tests of ESL reading comprehension. He does so through the use of a pretest-posttest study involving two groups of foreign students: the first group comprising 194 incoming students required to take the initial placement test, the second group, a subset of those, comprising 61 students who were placed into the reading course. Brown found that the item statistics indicated that a revision was possible and practical, that the revised version was effective as a norm-referenced reading placement test, and that the revised version was valid in terms of its construct validity. While a criticism of this paper is that a better assessment of the revised version of the test may come from an analysis of more than one administration of the test, and while Brown acknowledges that the revision "is just a beginning" (p. 79), the study nonetheless highlights the issue of accurate placement through Brown's devising a test that is closely related to what is taught in the program.

Another study which focuses on the issue of the importance of accurate placement of students in class levels is Rich (1993), who describes the addition of a writing sample to entry-

level testing. Rich tells of a college in Florida, where it was felt that the multiple-choice format of the entry-level placement test was not accurate in properly placing students into highly important college-preparatory English courses. Consequently, research was conducted to discover whether the addition of a writing sample would improve placements, how the addition could be accomplished in terms of practical matters (time, location, scoring, and so on), and whether the test-taker's score on a reading subtest might be useful for more accurate placement if the writing sample did not improve placement. In order to answer these questions, all students in what seems to be three levels of college-preparatory English classes were required to produce a writing sample, which was then sent to the College Board to be scored. In answer to the first question, the addition of a writing sample was found not to improve placements: "approximately 85% of all the writing samples scores indicated that students were properly placed. Very few students (29 of 1,399, or 2.1%) in the group received a judgment that they should placed in ESL . . . . Only 7% of the students received scores indicating a lower English course was needed" (p. 13). In answer to the second question, it was found that logistics were a problem: writing demands a place to write, a desk or table on which to write, and so on. In addition, writing creates essays to be marked, and it was determined that there was no way to mark the thousands of resultant essays; also, there was no determination as to who should mark them, and it was impossible to mark them without a week's turn-around time—an amount felt to be much too long. In answer to the third question, the scores on the reading and writing subtests were combined to test the effect of adding the reading placement score to the regression equation which already contained the writing placement score. The combination was found not to be entirely useful:

"In searching for a combination of reading and writing scores which could be used in placing students, one situation arose which made the combinations ineffective. In some cases where reading scores made significant contributions, the writing scores tended to cluster around the cut scores. This clustering was not consistently present. Thus before reading scores are placed in combination with writing scores, the cut scores for writing need to be reevaluated" (p. 16)

This study has its problems; for instance, the author mentions that at the college, "there are two levels of college preparatory course work in English, ENC0002 and ENC0020" (p. 2), yet she states that the writing test was administered to all students in ENC0002, ENC0020, *and* ENC1100, which is presumably either a third level of college preparatory course or a college course—whichever it is simply remains unexplained. Furthermore, Rich does not indicate how the writing topics were chosen nor who chose them, and she states that "the set of topics varied from campus to campus" (p. 4). This variation is a problem, as the topics may also vary in degree of difficulty. In spite of its problems, Rich's study nevertheless illustrates the issue of accurate placement through looking at crucial placement decisions at a college and the attempt to improve the placement vehicle through the addition of another type of test. Rich's study also demonstrates that what may seem to be a placement problem alone may in fact be a result of problems with the system that need to be addressed first.

Ilyin (1970) also notes the importance of accurate placement of students in class levels. She tells of the large numbers of students of varying abilities in each class of an adult ESL program in San Francisco and states that she "was surprised at the utter chaos that existed in our classes" (p. 1). After discovering that the program would have to develop its own placement

test, she describes the development of a standardized placement test to place the students into the first three levels of language classes and discusses work done on an experimental test to place students in the last three levels of language classes. In addition, she reports on a subsequent study to set norms and to investigate gains, but unfortunately, because the paper is outdated, the results are really of little use—after all, advances have been made in testing in the past three decades, and the use of discrete-point testing alone, as is suggested in Ilyin's paper, has been supplanted by other forms of testing (Weir, 1990). In spite of that, the paper underlines the importance of accurate placement. In Ilyin's words,

> we have one placement test for our lower levels that has been standardized and which has a high reliability . . . that place our students in classes better than previous methods. We still have to move a student or two, but not the large numbers of students we did before. The morale in the school is better. In short, both students and teachers are happier when placements are made more accurately." (p. 14)

Malu (1989), in her outline of ESL entrance testing and course placement procedures for the ESL program at the United Nations International School for grades 6-12, approaches the importance of accurate placement of students in class levels from a different perspective. Before doing so, however, she explains the procedure for determining whether a student needs to be tested for ESL; explains testing procedures, test, and interviews; and describes course placement and procedures. The difference is that instead of focusing on the placement test alone and its accuracy, Malu pinpoints the initial heavy investment of time taken during the placement test (as much as three hours in some cases) as having the beneficial result of "minimal class switching

because of misplacement" (p. 211). In terms of the thoroughness with which the program conducts its testing, placement decisions are not made on the basis of test results alone; rather, many other factors play a role, such as a holistic reading of an essay written as part of the placement test, and the student's background and behaviour during the test and the interview. While many other programs simply may not have the time to invest in a thorough assessment of their students during placement testing, and while Malu herself admits that "the major difficulty apparent to all who participate in this programme is the amount of time it takes to implement [the] procedures" (p. 211), Malu stresses the importance of accurate placement through a thorough testing procedure.

LeBlanc and Painchaud (1985), in their discussion of self-assessment as a placement instrument, also approach the importance of accurate placement from a different perspective. Rather than focusing on the refinement of a traditional testing method, they look at a technique typical of alternative assessment (as defined by O'Malley and Pierce, 1996): planned self-assessment. They maintain that accurate placements may be achieved through the use of questionnaires and that self-assessment testing can be a viable alternative to standardized testing. The authors describe a research project which aimed to answer the questions of whether students registering for second language courses at the university in which the authors worked could assess their own language proficiency, whether the type of self-assessment instrument used could influence the quality of this assessment, and whether self-assessment could be used as a placement instrument. To answer the first question, they randomly selected 200 students, who completed self-assessment questionnaires prior to taking the university's proficiency test. Afterwards, correlations were drawn between self-assessment and proficiency test scores, and it was found that students could indeed assess their own knowledge to some degree. To answer the

second, they gave two forms of the self-assessment questionnaire to students taking part in the fall registration at the university; one included metalinguistic vocabulary, the other did not. Correlations between the two forms were drawn, and it was found that the format seemed not to have any bearing on the quality of the answers. To answer the third, they revised level descriptions of second language courses, enlisted teachers to contribute "representative descriptors for each of the six levels in both listening and reading" (p. 683), and prepared a questionnaire based on the result. Afterwards, they tracked the percentage of level changes as the result of misplacement using the proficiency test in one academic year and the self-assessment test the next, and they found that "the self-assessment results placed the students at least as well as the standardized tests previously used" (p. 684). In fact, in all sessions, self-assessment seemed to have placed the students better than the standardized tests; level changes dropped between 1.5 to 3.7 percent from one year using the standardized tests to the next using the self-assessment placement test.

Although LeBlanc and Painchaud maintain that their students have the ability to assess themselves, and while all of their correlations are statistically significant at the .05 level, the correlations are, the authors admit, "not of the highest level" (p. 679), the one for the self-assessment speaking test, for example, being as low as .39. Also, the authors maintain that the format of the questionnaire did not matter as long as "students can understand the language used in the questions" (p. 682) and as long as the questionnaire was well-constructed. This particular group of students seemed to have little difficulty with the language of the questions in this case. However, new groups of students and their abilities may be different from those who come before them. What might happen if another group does have difficulty? Furthermore, while the authors state that the questionnaire was well-constructed, they give only a brief account of its

construction, never really explaining what they mean by "well-constructed." Finally, although level changes decreased by a few percent, these changes may have been made for reasons other than those related to misplacement—a fact the authors do admit—and the decrease itself was not examined to find out whether it is statistically significant. In spite of these shortcomings, LeBlanc and Painchaud suggest that self-assessment may be an accurate and valuable alternative to traditional placement instruments.

The Ethics and Effects of Testing

So far, testing has been observed from the point of view of the placement test itself: its definition, its component parts, its creation, and its accuracy in a variety of placement situations. No literature review would be complete, however, without a discussion of the ethics and effects of testing.

Shohamy (1993) argues that "few devices are as powerful, or are capable of dictating as many decisions, as tests" (p. 1). She maintains that

> results obtained from tests have serious consequences for individuals as well as for programs, since many crucial decisions are made on the basis of test results. Among these are the placement of students in class levels, the granting of certificates or diplomas, determinations as to whether students are capable of continuing in future studies, the selection of students most suitable for higher-education institutions, and the acceptance of job applicants and program candidates." (p. 2)

Shohamy holds that rather than providing information, tests have become tools for power and control, and she provides three examples in support of her argument: the first, the impact of the introduction of a test of Arabic as a second language; the second, the impact of a new EFL oral test; and the third, the impact of a reading comprehension test. Shohamy lists a number of findings, the most significant being those which follow. She finds that all three tests had some type of impact (as defined by Wall, 1997), and that the impact was complex and dependent upon the nature and purpose of the test. She notes, too, that the implementation of the tests caused instruction to become testlike, in other words there was "backwash," which is also known as "washback," (both terms as defined by Wall). She observes that the strength of the impact of these tests varied, depending on the type of test, subject relevance, the failure rate, and so on. Shohamy's paper is an important reminder that there is more to the creation of tests than just the simple assembly of test items: "tests are powerful devices and should be treated as such" (p. 18). Responsibility in test creation must begin somewhere, so "testers need to examine the uses that are made of the instruments they so innocently construct" (p. 19).

Responsibility in test creation could also be seen as a form of accountability. In reviewing and detailing the problems inherent in and development of accountability in language assessment, Norton (1997) points out that "language assessment practices should be accountable" (p. 313) not only to test-takers, who have been considered to be powerless stakeholders in the field of language assessment, but also to systems, because "schools, colleges and universities are under pressure to inform the public about what they are teaching and how effective they are" (p. 317). Accountability is an important consideration, and test-creators must think carefully about a number of matters: the uses to which their tests will be put; appropriate

training in testing and test use; and "recognition that test takers come from heterogeneous, culturally diverse backgrounds that must be taken seriously in the assessment process" (p. 314).

A good example of accountability is Peirce and Stein (1995), in which the authors describe their piloting a reading passage that was intended to be used as part of a South African college entrance examination for Black students. Because of concerns regarding the violence present in the then-current political climate, it was decided to pilot a reading passage to be given to Black students in a Johannesburg secondary school, for the reason that "if test takers became unduly disturbed by the content of the test, their performance might be compromised" (p. 54). It was found that the passage was interpreted as racist and was therefore rejected. Peirce and Stein's paper is a good example of accountability in that the authors were responsible to the test-takers and considering the possible results the test might have, piloted the passage before it was put into use.

STEPS TO TAKE IN THE CREATION OF A USEFUL AND ACCURATE PLACEMENT TEST

The preceding literature review has looked at several issues concerning placement testing. In "Testing and Assessment," the four types of test were discussed, and the components of language tests were examined. Language test creation, methodology, and practice were discussed, as well as steps and recommendations in test creation. In "Important Considerations," the significance of accurate placement testing and the steps that were taken to improve placements in various environments were examined, as were the ethics and effects of testing.

In this section, a series of steps will be reviewed to assist test creators in the writing of useful and accurate placement tests. The steps comprise the following: assembling an assessment team, defining the characteristics of the test-takers, defining the objectives for the placement test, deciding on the type of test to be used and its contents, creating the test, developing rubrics or rating scales for the test, testing the placement test itself, and training the scorers and administrators of the test.

Assembling an Assessment Team

Before the creation of a test can even begin, some preliminary steps are necessary. In discussing authentic assessments, O'Malley and Pierce (1996) suggest that the first step should be to assemble an assessment team. It matters not whether the assessment is to be authentic or otherwise: the idea of bringing together interested parties is an important one. This is the time to address individual stakeholders and their concerns on how to go about constructing the test. As Buck (1997) notes, "when designing tests, everything depends on the purpose of the test, and

the decisions that need to be made regarding the test-takers' ability. There will be advantages and disadvantages with any design, and compromises will usually be necessary" (p. 71). Assembling an assessment team should help the creator of a placement test to define the purpose of the test and to arrive at any decisions regarding the test. Concerning the composition of the assessment team, it should consist of any administrators who are responsible for curriculum and for students, of coordinators who are responsible for the implementation of the curriculum and the like, of teachers who represent a cross-section of the classes offered at the institute, and even of students at different levels within the school system. Noting also that test design involves compromises, Bradshaw (1990) states that "there seems to be no reason why some degree of collection of test-takers' and test-users' reactions cannot be included as part of the design of any new test" (p. 27).

Defining Characteristics of the Test-takers

After the objectives have been defined, the next preliminary step is to describe what type of test-taker will be taking the test. Carroll (1980) refers to this step as one of "participant identification" (p. 19) and includes it in the first of his recommended three phases of test construction. In identifying the test-taker, Carroll includes "relevant information about his identity and language background, such as his age, sex, nationality and place of residence as well as target language [and] mother tongue and any other languages learnt" (p. 19). Alderson, et al. (1995) include information on test-taker characteristics, such as age, gender, stage of learning, first language, cultural background, country of origin, type of education, reason for taking the test, personal and professional interests, and amount of background knowledge (p. 12). Having

access to this information should help the test creator greatly in choosing both appropriate material and test techniques.

Defining Objectives for the Placement Test

After the test-takers have been defined or characterized, the next preliminary step is to define objectives for the test. Hughes (1989) notes that this step is essential in testing "to make oneself perfectly clear about what it is one wants to know and for what purpose" (p. 48). Harrison (1983) believes that objectives for placement tests are different from those for other tests "because placement tests cannot be geared to the learning which went before" (p. 26). He suggests that test creators should think in terms of "aims," which Harrison says are more general than objectives. Semantics aside, the important point here is for test creators to decide what to test and how to do so. Although O'Malley and Pierce (1996) focus on authentic assessment, their belief that this step should encompass the determination of the purposes of the assessment and the specification of objectives is applicable really to any type of testing. In specifying the objectives, O'Malley and Pierce suggest—like Harrison—that objectives should be obtained from, among other sources, curricula.

Deciding on the Type of Test to be Used and Its Contents

Once the objectives for the test have been outlined, the next preliminary step is to decide what type of test is to be used and what to include in its contents. Concerning test type, three decisions must be made. First, should the test be direct or indirect, or a combination of the two. Second, should it be discrete point or integrative? Third, should it be norm- or criterion-referenced? Direct testing involves requiring the test-taker to perform the skill or skills to be

measured; indirect testing involves measuring the abilities underlying the skill. Hughes (1989) believes that while "it is preferable to concentrate on direct testing" (p. 16), he does admit that for some types of testing, indirect testing can be useful. Hughes observes that "direct testing is easier to carry out when it is intended to measure the productive skills of speaking and writing" (p. 15), and that indirect testing offers "the possibility of testing a representative sample of a finite number of abilities which underlie a potentially indefinitely large number of manifestations of them" (p. 16). With Hughes' observations in mind, for placement testing, it is recommended that a combination of the two approaches be used, with direct testing for speaking and writing, and indirect testing for listening and reading.

Discrete point testing involves testing one thing at a time, item by item; integrative testing involves testing the combination of many elements in the completion of a task. Hughes notes that the distinction between the two "is not unrelated to that between indirect and direct testing [and indeed] discrete point tests will almost always be indirect, while integrative tests will tend to be direct" (p. 17). Again, with Hughes in mind, it is recommended that for a grammar part (if included) and a listening part of a placement test, discrete point testing should be employed, while for a speaking part and a writing part, integrative testing should be employed. For reading, a combination of the two should be employed.

A norm-referenced test (NRT) is a test in which the amount of knowledge or material known by each test-taker is compared with that known by other test-takers, with the aim to spread students out along a continuum of general abilities or proficiencies so that differences among them are reflected in the scores (Brown, 1995). A criterion-referenced test (CRT) is a test in which the assessment of the amount of knowledge or material known by each test-taker is compared with a level of achievement or set of criteria. Subjectively marked tests are often

criterion-referenced (Alderson, et al., 1995). Brown (1995) says that since the purpose of CRTs "is to assess the amount of knowledge or material known by each individual student, the focus is on individuals rather than on distributions of scores" (p. 115). Brown also points out that, in contrast, the purpose of NRTs is to "generate scores that spread the students out along a continuum of general abilities or proficiencies in such a way that differences among the individuals are reflected in the scores" (p. 115). It is conceivable that on a CRT, all test-takers, if they know the material, could score 100 percent. While CRTs are mostly considered to be inappropriate as placement tests, which need to spread test-takers "out over a wide range of scores so that they can be sorted as efficiently as possible into class groups" (Harrison, 1983, p. 24), they have been used successfully in placement testing (Brown, 1989).

In any case, although Brown (1989) has successfully used a combination of NRT and CRT in what he has called a "new strategy for constructing language placements" (p. 73), NRTs are most often used for placement tests, and as such, are recommended for most placement situations. Nonetheless, if detailed criteria for classes or levels are already in place, CRTs certainly offer a viable alternative to NRTs in placement testing, and should be considered.

Concerning test content, it is logical that the placement test should reflect the curriculum of the school. In discussing the general development of language tests, Brown (1995) suggests that "a program-specific placement test could be developed so that the reasons for separating students into levels in the program are related to the things that the students can learn while in those levels" (p. 122). In practice, however, test items do not always mirror what is actually taught in class, and an example of this point is found in Brown's preamble to one of his earlier journal articles: "We decided to develop a placement battery that would be related in content to the curriculum of our institute—a proposal that struck us as strangely novel" (1989, p. 66).

A number of authors have proposed recommendations to assist test creators in deciding what to include in their tests. For example, Alderson, et al. (1995) recommend that at this stage, test creators ask themselves a variety of questions, such as how many sections the test should have, how long the sections should be, and how they should be differentiated; what the target situation is for the test and whether it should be simulated in some way; what text types should be chosen (written and/or spoken); what language skills should be tested; what language elements should be tested; what sort of tasks are required; how many items their should be in each section; and what test methods should be used. Chandavimol (1988) recommends that "the content of the placement test should directly reflect the parameters of the English programme concerned" (p. 3). Harrison (1983) advises both that the "contents of a placement test should be general" (p. 24) and that "the tests themselves should be fairly short, so that they do not take too long to answer or to mark" (p. 27). Most importantly, he recommends that "all four of the main language skills (listening, reading, writing, and speaking) should be tested" (p. 27).

In summary of these points, then, it is recommended that a NRT be used for a placement test and that depending on the context of school at which the test is to be employed, the test focus on the main language skills as they reflect the curriculum of the school.

Creating the Test

After the preceding preliminary steps, the first main step is to create the test itself. This step is important, for as Kirschner, Spector-Cohen, and Wexler (1996) indicate, "test questions constitute a communicative interchange between the test writer and the test taker" (p. 89). As such, then, the test creator must devise the test in such a way as to be "as easy for test takers to process as possible" (Kirschner, et al., p. 89).

Once this point has been understood, the test creator must then continue with creating the test and deciding on the parts of the placement test. As mentioned previously, listening, reading, writing, and speaking should be tested. In addition, a grammar component should be considered if the curriculum of the school places emphasis on grammar. In any case, in order to assist those who are tasked with the creation of a placement test, a brief set of general specifications—recommendations, really—for each part of a placement test follows these steps. Of course, depending on the testing context and other considerations, not every placement test will include all the recommendations listed here; as Brown (1995) warns, "many language tests are, or should be, situation specific" (p. 119). Nonetheless, for the sake of completeness, recommendations for each of the individual parts of a placement test have been included in this article, and it is suggested that those who are involved in the creation of placement tests use only those recommendations that apply to each individual testing situation.

Developing Rubrics or Rating Scales for the Test

The second main step is to develop rubrics or scoring guides for the placement test. Doing so should contribute to the reliable scoring of samples of the test-taker's performance. Although Harrison (1983) refers to rubrics in terms of "information for the student on how to do the test, including instructions, examples, and the organisation of test procedures" (p. 142), rubrics are taken here to refer to scoring scales that assign a numerical value to a test-taker's performance depending on the extent to which it meets pre-designated criteria (O'Malley & Pierce, 1996). As such, they are applicable to subjective or open-ended parts of the placement test, such as those containing short essay answers or oral interaction, and can be either holistic or analytical.

Holistic scoring "involves the assignment of a single score to a piece of writing on the basis of an overall impression of it" (Hughes, 1989, p. 86) and has the advantage of speed: Hughes notes that experienced scorers can assess a one-page piece of writing in "just a couple of minutes or less" (p. 86). One caveat concerning holistic scoring, however, is that the scoring scale must be very well conceived. Hughes points out that the rubric must "be appropriate to the level of the candidates and the purpose of the test" (p. 87). A second caveat is that there must be more than one scorer in order to ensure a high degree of scorer reliability. Analytical scoring requires "a separate score for each of a number of aspects of a task" (Hughes, p. 91) and has a variety of advantages. The most important of these are that scorers must consider certain aspects of the test-taker's performance that they might otherwise miss, that the results can be used for diagnostic purposes, and that "the very fact that the scorer has to give a number of scores will tend to make the scoring more reliable" (Hughes, p. 94). The main disadvantage with this type of rubric is that analytical scoring is time-consuming.

Which of the two types of rubrics should be developed by the creators of a placement test? Test creators must assess their testing situation and decide which to use. If time is at a premium, it is recommended that holistic scoring be used, for the reason that it is much more time and resource efficient, in that placement testing of objective items alone can be time-consuming, and testing is often done on-site with staff and/or faculty of the school in the role of test administrator and/or scorer. If there are enough time and resources, however, it is recommended that analytical scoring be used, for the reason that it can be the more reliable of the two and that the results can be used as a diagnostic tool by teachers of the classes into which the test-taker may be placed.

Testing the Placement Test Itself

The third main step is to analyze the newly created placement test, a step to which Alderson, et al. (1995) refer as pretesting and analysis. They state that "it is essential . . . that all tests should be pretested" (p. 74), because regardless of the care with which the placement test has been created, serious problems may exist with the test that cannot be identified during its conception. Harrison (1983) agrees, stating that "pretesting items is often regarded as essential . . . because trying them out with students shows how they work in practice, and it is only from this experimentation that bad items can be identified and amended or thrown out" (p. 127). Examples of problems with test items that may be identified at this stage are, according to Alderson, et al., (a) an abundance of items used in the test may be too difficult or too easy; (b) open-ended test items may confuse test-takers; (c) essay tasks may unintentionally result in less than adequate responses from the test-takers; and (d) multiple-choice items may be ambiguous and therefore open to disagreement (p. 74). Any one of these problems could serve to cause the placement test either to yield inaccurate results or not to work as intended—to spread students out on a continuum of language abilities.

Alderson, et al. thus suggest that the newly created test be trialed in exactly the same way as the final test will be on a number of students who are "representative of the final candidates, with a similar range of abilities and backgrounds" (p. 76). How many students are considered to be enough? As the authors point out, it is often difficult to find large numbers of students, so "the only guiding rule is 'the more the better,' since the more students there are, the less effect chance will have on the result" (p. 75). Once the test has been trialed, it should be analyzed. The authors suggest that objective test items, such as those of the multiple-choice variety, should be analyzed in terms of the facility value, which measures the level of difficulty of an item, and

the discrimination index, which measures "the extent to which the results of an individual item correlate with results form the whole test" (Alderson, et al., 1995, p. 80)[1]. The authors also suggest that subjective test items, such as those of the essay variety, should be analyzed in terms of "whether the items elicit the intended sample of language; whether the marking system . . . is usable; and whether the examiners are able to mark consistently" (p. 86).

Training the Scorers and Administrators of the Test

The last main step is to train the people who are going to be scoring and administering the placement test. As Underhill (1991) observes, "in testing, as in teaching, people are the biggest asset, and like any other resource, they can be used effectively or badly" (p. 15). Weir (1990) agrees that this step is important. He states that "considerable attention should . . . be paid to the development of relevant and adequate scoring criteria and examiners must be trained and standardised in the use of these" (p. 86). Alderson, et al. (1995) advise that

> the training of examiners is a crucial component of any testing programme, since if the marking of a test is not valid and reliable then all of the other work undertaken earlier to construct a 'quality' instrument will have been a waste of time. No matter how well a test's specifications reflect the goals of the institution or how much care has been taken in the design and pretesting of items, all the effort will have been in vain if the test users cannot have faith in the marks that the examiners give the candidates. (p. 105)

Alderson, et al. offer detailed advice concerning procedures for training the scorers of writing and speaking, and discuss the idea of having a Chief Examiner (p. 111) and standardization meetings (p. 112). While the creators of placement tests need not follow such a formalized

[1] For an explanation of how to calculate the facility value and/or discrimination index of test items, see, for example, Alderson, et al., pp. 80-86, Hughes (1989), pp. 161-162, or Harrison (1983), pp. 127-133.

method, it is nevertheless recommended that they develop a system that provides scorers with on-going training in the assessment of subjective test items.

Concerning administrators of tests, that is, those people who deliver the test to the test-takers, Alderson, et al. note that "though the training of administrators need not be as complex as that provided for examiners, it is still important that the administrators understand the nature of the test they will be conducting, the importance of their own role and the possible consequences for candidates if the administration is not carried out correctly" (p. 115). It is thus recommended that creators of placement tests also develop a system that provides administrators with training so that the test can be delivered consistently and correctly.

RECOMMENDATIONS FOCUSING ON INDIVIDUAL PARTS OF THE PLACEMENT TEST

Up to this point, this article has offered a detailed literature review and examined steps focusing on the creation of a placement test. A set of recommendations focusing on individual parts of the placement test now follows.

Listening

Buck (1997) observes that

> the basic idea of most listening tests is to assess the ability to use knowledge of the language for the purpose of understanding spoken texts . . . . [T]est tasks must [therefore] require fast, automatic, on-line processing of texts which have the typical linguistic characteristics of spoken language—especially the phonological characteristics. [In order to do so] non-interactive listening tasks . . . are probably most useful and certainly easier to construct. (p. 71)

To that end, Weir (1990) believes that the tasks should be authentic and "in terms of the tasks, items and scoring, it might be desirable in certain components of the test to focus on discrete items" (p. 52). The testing of listening should be accomplished, according to Harrison (1983), through the use of tape recordings, with the advantage being that the fact that the text is recorded makes it "more authentic, as if the students were actually listening to a radio talk or telephone message" (p. 29). Harrison conveniently omits the fact that not all listening tasks are conducted over the radio or telephone, but in spite of that, his point has merit for a very important reason: The test is more reliable because it is the same for each administration. As Harrison notes, "all students hear exactly the same text throughout all repeats and at all sittings of the test" (p. 29).

To summarize, the listening part of a placement test should seek to assess the ability of the test-taker to understand spoken language. The text should be authentic, and the tasks should

be non-interactive. The use of tape recordings is advised, and discrete-point testing is recommended, both for the sake of convenience of administration and marking, and for the sake of reliability.

Reading

Weir (1997) observes that a reading test "should reflect as closely as possible the interaction that takes place between a reader and a text in the equivalent real life reading activity" (p. 39). Therefore, the approach to the reading part of a placement test should be direct. Hughes (1989) believes that there are at least four levels of reading that can be tested: low-level operations, grammatical and lexical abilities, macro-skills, and micro-skills. The term "low-level operations" refers, for example, to the ability to distinguish between letters of the alphabet, e.g., between "b" and "d." According to Hughes, there is no call for the formal testing of this ability in that information on this ability can be observed through informal observation. Grammatical and lexical abilities refer to the ability, for example, to use the present perfect aspect or to define vocabulary. Information on these abilities can be collected, as Hughes notes, "through tests of grammar and vocabulary, not necessarily as an integral part of a reading test" (p. 117). Macro-skills refers to the ability to scan text to find specific information, to skim to obtain gist, to identify the support of an argument, and so on, while micro-skills refers to the ability to identify referents of pronouns, to use context to guess meanings, to understand transition words, and so on. While a test of macro-skills is possible, Hughes believes that "only at the level of . . . 'micro-skills' do we reach the point where we find serious candidates for inclusion in a reading test" (p. 117).

The text of the reading part of a placement test does not necessarily have to be authentic—a term that Hughes defines as "intended for native speakers" (p. 118). Instead, Hughes suggests that whether or not authentic texts are employed in any sort of reading test depends in part on what the test is intended to measure. Unfortunately, he does not offer any further information on what type of measurement demands what kind of text (authentic or not), but he does state that "even at lower levels of ability, with appropriate items, it is possible to use authentic texts" (p. 118). In light of that point, then, authentic texts are recommended, the types of which, according to Hughes, might include textbooks, novels, magazines, newspapers, journals, and timetables—to name a few. The type may be further specified, such as a two- or three-paragraph passage from a novel, an article from a magazine, or an advertisement in a newspaper.

A number of techniques can be employed in the testing of reading, but Hughes cautions that "we have to recognise that the act of reading does not demonstrate its successful performance. We need to set tasks which will involve candidates in providing evidence of successful reading" (p. 120). The difficulty is, however, employing techniques or tasks which do so without interfering in the reading itself. Hughes offers a list of techniques, including multiple-choice, short answer, guided short answer, and information transfer. While the disadvantages of multiple-choice testing are well-documented (in, for example, Hughes, 1989; Weir, 1990; Weir, 1997), it should be noted that multiple-choice testing is reliable and does lend itself well to rapid scoring. Short answer and guided short answer testing may provide a good indication of reading ability, but both techniques have the disadvantage of the potential for obscuring the test-taker's true ability because each demands the ability to write: to use Hughes example, "a student who has the answer in his or her head after reading the relevant part of the

passage may not be able to express it well" (p. 122). Information transfer, on the other hand, has the advantage of minimizing the potential for obscuring the test-taker's ability in that this technique demands little or no writing ability.

To summarize, the reading part of a placement test should seek to assess the ability of the test-taker to understand written language. The text should be direct and authentic. A variety of techniques are recommended: multiple-choice for its ease of scoring, short answer or guided short answer for its indication of ability, and information transfer for its lack of dependence on the test-taker's ability to write. The choice of which technique to employ is a difficult one, and must be decided according to each individual placement testing situation. In light of that point, a combination of the above techniques is recommended.

Speaking

Underhill (1991) observes that "when we test a person's ability to perform in a foreign language, we want to know how well they can communicate with other people, not with an artificially-constructed object called a language test" (p. 5). In considering a test of speaking, then, Weir (1990) believes that "the essential task for the test designer is to establish clearly what activities the candidate is expected to perform, how far the dynamic communicative characteristics associated with these activities can be incorporated into the test, and what the task dimensions will be in terms of the complexity, size, referential and functional range of the discourse to be processed or produced" (p. 74). Underhill (1991) takes a more humanistic approach to the testing of speaking by stating that "oral tests must be designed around the people who are going to be involved. This is a human approach; we want to encourage people to talk to each other as naturally as possible. The people, not the test instrument, are our first concern"

(p. 4). To underscore that point, Underhill continues by suggesting that "the direct interview is the most common and most authentic type of oral test for normal purposes; there is no script and no preparation on the learner's part for any special activity" (p. 31).

Regardless of the type of speaking test, Weir cautions that "in oral testing . . . there is a need for explicit, comprehensive marking schemes, close moderation of test tasks and mark schemes, and rigorous training and standardisation of markers in order to boost test reliability" (p. 80). On this point, Underhill adds that accurately-worded rating scales will be of great benefit to those administering the speaking test (p. 13). In addition, Alderson, et al. (1995) suggest that the person administering the test of speaking is important "because it is always necessary for at least one person to elicit language from the candidate and to react in an encouraging way to keep the language flowing" (p. 116). Underhill echoes this point: "the interviewer should also know a lot about what happens in [the] classes. Ideally, she should be a regular class teacher herself so she knows the classes well and can ask herself questions like, 'How would I feel if this learner appeared in my class tomorrow?'" (p. 13). Also of importance, according to Alderson, et al., is an environment that will not be intimidating to the test-takers, one "which will help candidates to feel at ease" (p. 117).

To summarize, the speaking part of a placement test should seek to assess the ability of the test-taker to communicate orally with other people, not with a testing instrument. A direct interview is recommended, as it is not scripted and requires no special preparation on the part of the test-taker. However, comprehensive marking schemes should be devised, and precisely-worded rating scales are recommended. The person administering the speaking test should be a regular class teacher, and the speaking test should be administered in a place that is comfortable to the test-taker.

Writing

Hughes (1989) assumes that "the best way to test people's writing ability is to get them to write" (p. 75); thus, the approach to the writing part of a placement test should be direct. The tasks contained in the writing test should, according to Hughes, be "representative of the population of tasks that we should expect the students to be able to perform [and] should elicit samples of writing which truly represent the students' ability" (p. 75), and do not represent other things such as the creativity, imagination, or intelligence of the test-taker. In addition, Hughes maintains that the samples of writing obtained from these tests should be scored reliably. While a large number of writing tasks is seen by Hughes as being ideal in terms of validity (p. 81), it is impractical in a placement test. As Hughes notes, "if it is a matter of placing students in classes from which they can easily be moved to another more appropriate one, then accuracy is not so important; we may be satisfied with a single sample of writing" (p. 82).

While a number of strategies exist for testing writing ability, Weir (1990) offers three viable suggestions: the summary task, the controlled writing task, and the essay test. Summarizing, however, is problematic in that it demands the production of a specific text, one which might be too narrow and thus beyond the knowledge or abilities of the test-taker. The controlled writing task, while necessary "where writing tasks are an important feature of the student's real life needs" (Weir, p. 61), is also problematic in that there may be situations "when the complexity of the stimulus obstructs the desired result, i.e., one needs to understand a very complex set of instructions and/or visual stimuli to produce a relatively straightforward description of a process or a classification of data" (p. 62). The essay test is problematic in that it is open-ended, timed, and time-consuming. In addition, among other problems associated with

this form of testing, the ability to write freely on topics "may depend on the candidate's background or cultural knowledge, imagination, or creativity" (p. 60). Nevertheless, in spite of the problems associated with the essay test, it is a traditional method for testing writing ability, is familiar to a wide variety of test-takers, and thus holds much face validity.

To summarize, the writing part of a placement test should seek to assess the ability of the test-taker to write. A direct test is advised, as testing writing through the use of indirect, discrete-point items does not clearly give an indication of writing ability (Weir, 1990, p. 59). The essay test is recommended as the vehicle for testing writing ability, yet caution must be taken in the creation of topics for the essay: test-takers may be hampered in that the topic may be uninteresting or culturally biased. It is therefore recommended that a selection of topics be offered on a variety of subjects, with the test-taker writing on one topic.

Grammar

Rea-Dickens (1997) notes that the communicative approach to language teaching has lessened "the role of grammar as a respectable focus of teaching and learning" (p. 94), yet Hughes (1989) observes that

> there is often good cause to include a grammar component in the achievement, placement and diagnostic tests of teaching institutions. It seems unlikely that there are many institutions, however 'communicative' their approach, that do not teach some grammar in some guise or other. (p. 142)

While the testing of grammar has traditionally been accomplished through the use of multiple-choice items, other techniques are available and may even be preferable; rather than requiring the test-taker solely to recognize correct use, as is the case in most multiple-choice items, these techniques require that the test-taker use grammatical structures appropriately. Hughes lists three such techniques: paraphrase, completion, and modified cloze. Paraphrase requires the test-

taker to write a sentence, the beginning of which is supplied, that is similar in meaning to one that is given. Completion requires the test-taker to complete sentences by supplying correct structures in context (for example, interrogative forms in the completion of questions, with the responses already supplied). Modified cloze requires the test-taker to complete sentences by supplying the deleted form (for example, prepositions or articles).

To summarize, because grammar is taught in some way or other, it should be considered as a component of a placement test. The multiple-choice and modified cloze techniques are recommended for their ease of scoring.

CONCLUSION

Placement testing is important, yet it can be problematic. With thought and careful planning, however, problems can be minimized, and placement testing can be more accurate.

REFERENCES

Alderson, J. C., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation*. Cambridge: Cambridge University Press.

Bradshaw, J. (1990). Test-takers' reactions to a placement test. *Language Testing, 7*(1), 13-30.

Brown, J. D. (1989). Improving ESL placement tests using two perspectives. *TESOL Quarterly, 23*(1), 65-83.

Brown, J. D. (1995). *The elements of language curriculum: A systematic approach to program development*. New York: Heinle & Heinle.

Buck, G. (1997). The testing of listening in a second language. In C. Clapham and D. Corson (eds.), *Encyclopedia of language and education, volume 7: Language testing and assessment* (pp. 65-74). Netherlands: Kluwer Academic Publishers.

Carroll, J. C. (1980). *Testing communicative performance: An interim study*. Oxford: Pergammon Press.

Chandavimol, M. (1988). *The placement test: A useful or harmful tool.* (ERIC Document Reproduction Service No. ED 298 769)

Courchêne, R. (1995). An alternative method for teaching and testing reading comprehension. *TESL Canada Journal, 12*(2), 50-58.

Cumming, A. (1997). The testing of writing in a second language. In C. Clapham and D. Corson (eds.), *Encyclopedia of language and education, volume 7: Language testing and assessment* (pp. 51-63). Netherlands: Kluwer Academic Publishers.

Elson, N. (1992). The failure of tests: Language tests and post-secondary admission of ESL Students. In B. Burnaby and A. Cumming (Eds.), *Socio-political aspects of ESL* (pp. 110-121). Toronto: OISE.

Fulcher, G. (1997). The testing of L2 speaking. In C. Clapham and D. Corson (eds.), *Encyclopedia of language and education, volume 7: Language testing and assessment* (pp. 75-85). Netherlands: Kluwer Academic Publishers.

Harrison, A. (1983). *A language testing handbook*. London: Macmillan Press.

Hughes, A. (1989). *Testing for language teachers*. Cambridge: Cambridge University Press.

Ilyin, D. (1970). Developing a placement test for adults in English-second-language programs in California. (ERIC Document Reproduction Service No. ED 036 766)

Kirschner, M., Spector-Cohen, E., & Wexler, C. (1996). A teacher education workshop on the construction of EFL tests and materials. *TESOL Quarterly, 30*(1), 85-111.

LeBlanc, R. & Painchaud, G. (1985). Self-assessment as a second language placement instrument. *TESOL Quarterly, 19*(4), 673-687.

Lynch, B. K. and Davidson, F. (1994). Criterion-referenced language test development: Linking curricula, teachers, and tests. *TESOL Quarterly, 28*(4), 727-743.

Malu, K. F. (1988). Entrance testing and course placement at the U.N. International School, New York City. *ELT Journal, 43*(3), 206-212.

O'Malley, J. M. & Pierce, L. V. (1996). *Authentic assessment for English language learners: Practical approaches for teachers*. Addison-Wesley.

Peirce, B. N. & Stein, P. (1995). Why the "Monkeys Passage" bombed: Tests, genres, and teaching. *Harvard Educational Review, 65*(1), 50-65.

Progosh, D. (1996). Using video for listening assessment: Opinions of test-takers. *TESL Canada Journal, 14*(1), 34-44.

Rea-Dickens, P. (1997). The testing of grammar in a second language. In C. Clapham and D. Corson (eds.), *Encyclopedia of language and education, volume 7: Language testing and assessment* (pp. 87-97). Netherlands: Kluwer Academic Publishers.

Read, J. (1997). Assessing vocabulary in a second language. In C. Clapham and D. Corson (eds.), *Encyclopedia of language and education, volume 7: Language testing and assessment* (pp. 99-107). Netherlands: Kluwer Academic Publishers.

Rich, J. C. (1993). Can a writing sample improve placement in English courses? Research Report No. 93-13R. (ERIC Document Reproduction Service No. ED 366 400)

Shohamy, E. (1993). The power of tests: The impact of language tests on teaching and learning. (ERIC Document Reproduction Service No. ED 362 040)

Underhill, N. (1991). *Testing spoken language: A handbook of oral testing techniques.* Cambridge: Cambridge University Press.

Weir, C. J. (1990). *Communicative language testing.* New York: Prentice Hall.

Weir, C. J. (1997). The testing of reading in a second language. In C. Clapham and D. Corson (eds.), *Encyclopedia of language and education, volume 7: Language testing and assessment* (pp. 39-49). Netherlands: Kluwer Academic Publishers.

About the Author

Joel Murray (MA TESL, University of British Columbia), whose main interest is language testing, has been a full-time ESL instructor since 1981. In his many years of teaching, he has taught a variety of levels of ESL, from beginner with no prior exposure to English to highly advanced. He currently specializes in teaching academic preparation courses for ESL students who are planning to enroll in, or who are currently enrolled in college programs or university transfer courses at Kwantlen University College, Canada's largest public university college, serving Vancouver and its suburbs.

U.S. Department of Education
Office of Educational Research and Improvement
(OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

Reproduction Release

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: Steps and Recommendations for More Accurate Placement Test Creation	
Author(s): Joel R. Murray	
Corporate Source:	Publication Date: May, 2001

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, Resources in Education (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign in the indicated space following.

The sample sticker shown below will be affixed to all Level 1 documents	The sample sticker shown below will be affixed to all Level 2A documents	The sample sticker shown below will be affixed to all Level 2B documents
PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY SAMPLE TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY SAMPLE TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY SAMPLE TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
Level 1	Level 2A	Level 2B
☑	☐	☐
Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g. electronic) *and* paper copy.	Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only	Check here for Level 2B release, permitting reproducti and dissemination in microfiche only
Documents will be processed as indicated provided reproduction quality permits. If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.		

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche, or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Signature:	Printed Name/Position/Title: Joel R. Murray	
Organization/Address: Kwantlen University College 12666 - 72nd Ave. Surrey, BC V3W 2M8 CANADA	Telephone: 604-599-2787	Fax: 604-599-2716
	E-mail Address: joel@kwantlen.bc.ca	Date: May 22, 2001

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, or, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:
Address:
Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:
Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

ERIC Processing and Reference Facility
4483-A Forbes Boulevard
Lanham, Maryland 20706
Telephone: 301-552-4200
Toll Free: 800-799-3742
e-mail: ericfac@inet.ed.gov
WWW: http://ericfac.piccard.csc.com

EFF-088 (Rev. 9/97)